



## Call for Applications



### History as a Data Science

#### Free Two-Week Research Workshop at Columbia's History Lab

Digital history is thriving, and there are many resources available for using it in the classroom. But the increasing volume of digitized and “born digital” collections of books, articles, and archives also presents tremendous new opportunities for original research. This is especially true in fields like international history, where the declassification of official papers in online databases has created some of the largest collections of historical documents in the public domain. But for those with little or no training in data science, the only way to explore these resources has been through keyword searching.

The goal of this workshop is to offer a very practical introduction to the many other methods of using large digitized archives only possible with direct access to the data. Participants will learn how to organize and analyze textual data and get an overview of advances in natural language processing and machine learning. Hands-on training will use textual data from [History Lab](#), an NSF-funded project that has aggregated the largest database of declassified government documents in the world. Participants will also learn how to get their own data by “scraping” websites and downloading from online databases. More specifically, we will examine how to bring textual data into Python and R, how to use Python for web scraping, and how to explore textual data using string functions. These methods make it possible to grapple with old research problems with new rigor, and launch entirely new kinds of inquiries.

The workshop is timed to follow immediately after the 2020 American Historical Association meeting in New York City. Those interested in applying are encouraged to attend the AHA and participate in the digital history panels and workshops. Thanks to support from Columbia University and a grant from the American Council of Learned Societies, we have funds to subsidize the travel and lodging costs of out-of-town participants. But we also ask you to apply for your own funds so we can make this opportunity available to others.

**When:** January 6, 2020 – January 17, 2020. Sessions will be from 9am - 1pm each weekday, with individual and small-group meetings to follow. If you would like to participate but the timing poses a problem, tell us which dates and times you can commit to attending.

**Where:** Columbia University in NYC, at the Institute for Social and Economic Research and Policy.

**Eligibility:** This workshop is open rank: first year Ph.D. students through established scholars are encouraged to apply. Priority will be given to historians. Others will be eligible to participate on a space available basis.

#### Course Objectives

At the end of the course, participants should be comfortable with the basics of digitized text. This includes being able to open text data from a variety of sources and understand its formatting and structure; knowing what can and cannot be done with digital data; and developing experience in some fundamental analysis techniques.

## Course Prerequisites

Experience with coding and familiarity with different computer languages will be extremely helpful but is not required. Participants with no background will be encouraged to do some preparatory work.

## Instructors

**Matthew Connelly** is a professor of history at Columbia and the principal investigator of History Lab. He received his B.A. from Columbia and his Ph.D. from Yale. His publications include *A Diplomatic Revolution: Algeria's Fight for Independence and the Origins of the Post-Cold War Era*, which won five prizes, and *Fatal Misconception: The Struggle to Control World Population*, an Economist and Financial Times book of the year. He has provided commentary on international affairs for The Atlantic Monthly, The New York Times, The Washington Post, and Le Monde, and has hosted radio documentaries for BBC Radio.

**Raymond Hicks** is a research scholar at History Lab and has a Ph.D. in Political Science from Emory. Before starting at Columbia, he worked as the Statistical Programmer for the Niehaus Center for Globalization and Governance at Princeton University. His research interests include monetary policy, trade policy, and statistics and his work has appeared in the *Journal of Politics*, *International Organization*, and the *British Journal of Political Science*, among other journals.

Additional presentations will be given by leading data scientists and digital historians.

## How to Apply

Please submit your application to Harlowe Wang, the workshop coordinator, at [hw2713@columbia.edu](mailto:hw2713@columbia.edu). Applications will be considered on a rolling basis. Please include the following materials:

- 1-2 page letter of application. The letter should include:
  - What background (if any) you have in digital methods
  - What area of research you could imagine applying such methods. If you have a specific idea for a project, please include an abstract.
  - What you hope to learn from the workshop, and how it fits into your career plans.
- CV
- Contact information for your advisor OR one other reference (only required for graduate students)
- Budget with your costs for attending (i.e. travel and lodging) and any other funding sources to which you are applying. Affordable options near Columbia University can be found on AirBnB (from about \$40/night), and we will encourage participants to share multi-room apartments.

Please email Harlowe Wang, the workshop coordinator, at [hw2713@columbia.edu](mailto:hw2713@columbia.edu) with any questions regarding the workshop or application process.

This free workshop is sponsored by the American Council of Learned Societies.

